**Заявленный результат - Разработка онтологии и технологии унифицированного доступа к данным результатов исследований на основе Linked-Data и современных стандартов представления данных о биомах. Интеграция хранилища данных проекта в мировую информационную среду.**

**Цифровой архив размеченных документов – ядро сервиса интеграции в мировую информационную среду исследований в биологии**

Целью разработки данного направления является создание средств интеграции результатов исследований в проекте с данными других исследований, проводимых в России и за рубежом. Такая интеграция позволит частично автоматизировать обработку информации, содержащейся в материалах, первоначально предназначенных для ознакомления человеку, например, отчетов, статей, страниц Интернет, баз данных и т.д.

Исследования велись по следующим направлениям:

1. Поиск онтологий для представления публикуемых документов.
2. Разработка системы хранения и обеспечения публикуемых документов в частично формализованном виде.
3. Создание средств обеспечения доступа к документам, включая поиск по содержимому.
4. Разработка программной системы для построения логического вывода на данных документов.

В основе концепции публикации данных является поддержка технологий открытых связанных данных (Linked Open Data). Технологии позволяют осуществлять интеграцию данных на уровне глобальных идентификаторов ресурсов (URI, Universal Resource Identifiers) и общих стандартных концептуальных моделей (онтологий), используемых для представления данных (атрибутов и связей) публикуемых ресурсов.

Как правило, в процессе публикации ресурсов в Интернет решаются следующие технические проблемы:

1. Представление структуры публикуемых объектов при помощи стандартных форматов, позволяющих предоставлять доступ к имеющейся информации внешним ресурсам, алгоритмически обрабатывать публикуемую информацию.
2. Реализация серверов непосредственно предоставляющих требуемую информацию по запросу: по идентификатору или характеристике содержимого, например, по ключевым словам, тегам или части текста.
3. Создание средств обнаружения публикуемой информации: регистрация собственных ресурсов в известных системах поиска информации.

На этапе исследований по проекту в 2017 году удалось продвинуться в решении первых двух задач: изучен и проанализирован ряд стандартных онтологий, используемых для публикации данных разного рода, в том числе научных данных. Разработано программное обеспечение для непосредственного предоставления ресурсов по запросу (цифровой архив), поддерживающий несколько видов запросов (по содержимому, по адресу ресурса и SPARQL). Цифровой архив снабжен средствами логического вывода для поддержки решения задач связанных с предварительной обработкой запрашиваемой информации на основе формализованных знаний. Процедуры логической обработки также запускаются в результате соответствующего типа запроса (Pengines), в результате чего порождаются новые табличные данные – качественные суждения о наборе объектов.

Основным публикуемым документом в проекте является научная статья. Статья включает содержательную информацию, представляемую в виде текста, таблиц, графиков и рисунков. На первом этапе рассмотрена и решена задача структурной разметки документа, организацию документов в структуры зависимостей по передаче информации. Полученные результаты позволяют нацеливаться на решение задач разметки текста, представляющего закономерности, логические и ассоциативные связи между объектами и явлениями, реализуемых на следующих этапах.

Структурная разметка материалов публикаций использует следующие онтологии:

1. [Open Annotation](http://www.openannotation.org/) (oa). Основная задача данной онтологии - представлять содержимое (аннотацию), описывающее другое содержимое. Закладка браузера является примером такого содержимого.
2. [Friend-of-a-friend](http://xmlns.com/foaf/spec/) (foaf) позволяет представлять информацию об агентах: физических и юридических лицах, а также программных агентов.
3. [Provenance](https://www.w3.org/TR/prov-o/) (prov). Онтология prov – основа описания информационных потоков в документах и их взаимосвязи. В цифровом архиве prov используется для ассоциации документа с цифровым архивом и организацией владельцем.
4. [Dublin Core](http://dublincore.org/documents/dcmi-terms/) (dc) представляет в аннотации ментаинформацию о творческом произведении: авторов, формат содержимого, его описание и др.
5. [DBPedia](http://wiki.dbpedia.org/) resource (dbr) – пространство имен объектов (ресурсов) Wikipedia. Онтология dbr используется для обозначения конкретных сущностей, географических названий и т.п.
6. [Bibliographic ontology](http://bibliographic-ontology.org/specification) (bibo) предназначена для разметки ссылок на внешние публикации, активно использует онтологию dc.
7. [Schema.org](http://schema.org/) (schema) представляет объекты, распознаваемые поисковыми агентами Google, Yandex, Yahoo и др.

Кроме перечисленных онтологий используются стандартные онтологии (RDF, RDFs, RDFa, XSD) и онтологии из проекта [NEPOMUK](https://userbase.kde.org/Nepomuk), предназначенные для описания объектов, хранимых в полнотекстовых индексных цифровых архивах.

Предметная область NGS отражена в следующих онтологиях:

1. [Semanticscience Integrated Ontology](https://github.com/micheldumontier/semanticscience) (sio) задает простую интегрированную модель для объектов, процессов и их атрибутов. Онтология обеспечивает фундамент известных проектов по интеграции биологических данных [Bio2RDF](http://bio2rdf.org) и [SADI](http://sadiframework.org) ([Semantic Automated Discovery and Integration](http://sadiframework.org/content/about-sadi/)).
2. [TaxonMap Ontology](http://purl.org/biodiversity/taxon/) (taxon) - словарь для отображения таксономических классов в облако открытых связанных данных.
3. [Biological Taxonomy Vocabulary](http://purl.org/NET/biol/ns) (biol) - словарь для представления таксономии всех форм жизни.
4. [BioPAX Level 3 ontology](http://www.biopax.org/) (biopax) - онтология для представления процессов преобразования веществ и влияний (катализ, ингибирование и т.п.) веществ на биологические процессы. Она позволяет ученым обмениваться информацией друг с другом. Применение данной онтологии позволяет уменьшить сложность представления информации.
5. [BioTop](http://biotopontology.github.io/) (biotop) используется для представления функциональных сущностей в биологии и медицине, представляет собой онтологию верхнего уровня, совместимую с онтологиями BFO, DOLCE, и UMLS Semantic Network.
6. [Ontology for Biomedical Investigation](https://raw.githubusercontent.com/obi-ontology/obi/v2017-09-03/obi.owl) (obi) предназначена для разметки биомедицинские исследования, включая план исследования, протоколы и использованные приборы, а также позволяет представлять типы проводимых анализов и получаемые данные. Онтология obi получена усовершенствованием онтологии "Functional Genomics Investigation Ontology" (FuGO). Дальнейшее ее развитие направлено на включение концепций функциональной геномики и связанных с ней предметных областей.

Популярным подходом к решению проблемы обнаружения данных является регистрация сайтов/порталов в поисковых ресурсах, таких как Google или Yandex. Данные информационные системы позволяют эффективно обнаруживать ресурсы Интернета (конкретные статические и динамические страницы) по ключевым словам и отрывкам текста. Для серверов, где все публикуемые ресурсы (страницы) являются динамическими и генерируются во время запроса их URI, данные сервера не предоставляют эффективной процедуры обнаружения ресурсов по контексту, т.е. отображение свойств содержимого на URI ресурса. Повышение эффективности обнаружения и предоставление информации об общей структуре ресурса используются специализированные средства, такие как Bio2RDF.

Проект Bio2RDF направлен разработку технологий обеспечения среды связанных данных в науках о жизни. В основе разрабатываемых технологий находится так называемый Семантический веб, программы и модули для обработки соответствующим образом размеченной информации. Bio2RDF определяет набор простых форматов данных и принципов их обработки, создающих возможности [интегрирования данных из разных источников](https://docs.google.com/presentation/d/1SG6PFew2CPK1o_jRCYx30DFnuGEEr0uWVKPnz7Uqw5k/pub?start=false&loop=false&delayms=3000&slide=id.p), представленных в различных форматах в рамках технологий связанных данных. Действующий сервер, сопровождаемый разработчиками, демонстрирует принципы Bio2RDF и семантические технологии. Доступный исходный код позволяет свободно создавать собственные ресурсы.

Сервисы Bio2RDF активно используют вышеперечисленные онтологии для представления данных. Например, prov используется для указания источника полученных данных. Сервер хранит ссылки на ресурсы источника данных в виде отношений идентификаторов объектов в нотации ресурса и нотации Bio2RDF. Кроме того, для всех объектов должен быть присвоен тип (класс). Таким образом, сервисы Bio2RDF позволяют организовывать несколько важнейших функций публикации данных – обнаружение данных в результате выполнения запросов на центральном сервере метаданных, обеспечение ссылок на ресурсы в виде стандартных URL, что в свою очередь позволяет загружать предоставленные данные и адаптировать их форматы к собственным алгоритмам. Использование языка SPARQL на сервере метаданных позволяет включать ресурсы сервера в комплексные запросы, задействующие несколько серверов семантических данных. В результате запроса, как правило, порождается ответ, похожий на реляционную таблицу. Такой подход и обеспечивает интеграцию данных из разрозненных источников.

Публикация ресурсов осуществляется при помощи отображения URI ресурса на его содержимое при помощи серверов, функционирующих по стандартному протоколу HTTP. Реализация таких серверов обычно представляет собой цифровой архив размеченных документов. В рамках проекта разработаны архитектура и подсистемы такого цифрового архива, поддерживающего технологии открытых связанных данных (рис. [1](file:///\\home\eugeneai\Development\text\Projects\Lin-2017\report-2017\annotation.html#fig:architecture-LOD)).

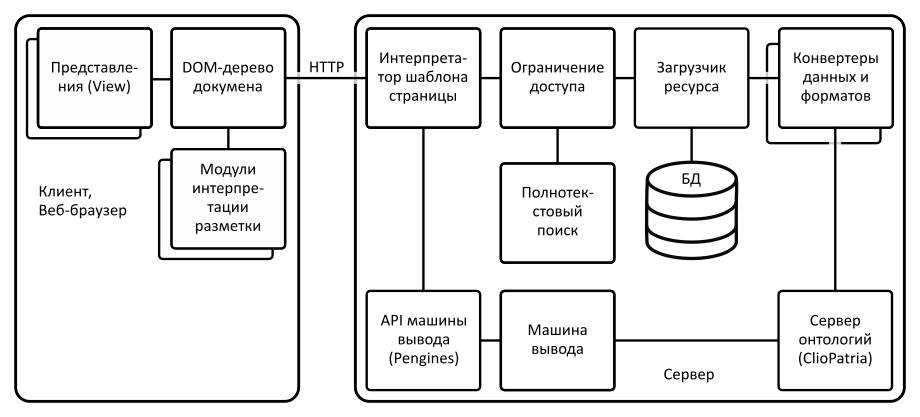


Рис. 1. Цифровой архив ресурсов проекта с обеспечением поддержки открытых связанных данных

Цифровые данные содержимого ресурсов хранятся в базе данных системы. Система поддерживает множество таких баз данных. База данных взаимодействует с загрузчиком данных, задача которого состоит в преобразовании URI ресурса в содержимое. Для каждого источника необходимо разрабатывать специальные конвертеры, задача которых состоит в преобразовании публикуемой информации из базы данных в виде содержимого, размеченного в соответствии с принципами открытых связанных данных. В качестве форматов представления содержимого ресурсов выступают размеченные RDFa HTML5-страницы или документы, а также графы онтологий, представленные в TTL-, RDF- и OWL2-форматах.

Часть документов храниться в системе в “разобранном” виде и собирается в веб-браузере в готовую статью. Модули интерпретации (порождения текста документа) семантической разметки таких документов реализованы при помощи клиентского JavaScript-интерпретатора, встроенного в веб-браузер. Модули сканируют структуру дерева документа, распознавая условия их активации. Если поиск оказался успешным, запускается тело модуля, которое в общем случае вносит изменения в документ. Формирование документа заканчивается, как только все условия были активированы и все команды были обработаны.

Модуль разграничения доступа позволяет предоставлять доступ к данным на основе некоторой политики, например, в соответствии с правами зарегистрированного пользователя. Система позволяет загружать только те ресурсы, т.е. получать доступ к URI ресурсов, соответствующие полномочиям пользователя. Сервис полнотекстового поиска позволяет ассоциировать тексты содержимого ресурсов с URI этих ресурсов. Подсистема поиска реализована при помощи библиотеки Elasticsearch, которая обладает средствами нечеткого сравнения термов, что позволяет развивать систему в направлении реализации систем поиска релевантной информации.

Интеграция с сервисами обнаружения ресурсов имеет несколько уровней взаимодействия. На самом простом уровне сервер обнаружения лишь предоставляет запрашивающей стороне только адреса URI зарегистрированных ресурсов. Более тесная интеграция позволяет серверам выполнять совместные (федеративные) запросы SPARQL. Для создания такой возможности в цифровой архив добавлены блоки поддержки SPARQL, включая базу данных и знаний (сервер онтологий), который используется для хранения предварительно подготовленной логической структуры публикуемого ресурса. Сервер онтологий реализован при помощи программной системы ClioPatria. Данная система интересна тем, что полностью реализована на языках Prolog и С, она предоставляет несколько форматов компактного хранения данных, а также тесную интеграцию данных открытых связанных данных со средой программирования языка Prolog.

Машина логического вывода языка программирования Prolog используется для предварительной обработки публикуемой информации, а также для обеспечения доступа к знаниям системы извне. Такой доступ обеспечивается при помощи "API машины вывода" (Pengines). Данный интерфейс (API) позволяет разрабатывать HTML5-JavaScript-приложения, отображающие тем или иным способом данные из цифрового архива.

Благодаря использованию языка программирования Python и технологий открытых связанных данных, а также модулей интерпретации разметки, разработанный сервис публикации хорошо интегрируется с базами данных, в том числе с разрабатываемой в рамках проекта базой данных научных исследований.

Таким образом, на этапе 2017 года по данному направлению решены две задачи – разработан подход к описанию публикуемых данных при помощи набора стандартных онтологий, разработан сервис публикации размеченных документов, хорошо интегрирующийся с существующими цифровыми архивами и базами данных.

**Публикации по направлению**

1. Черкашин Е.А., Шигаров А.О., Орлова И.В., Михайлов И.С. Использование технологий Linked Open Data при подготовке и публикации текстовых документов. // Материалы Всероссийской конференции с международным участием "Знания-Онтологии-Теории" (ЗОНТ-2017), 2-6 октября 2017 г., Новосибирск, Институт математики им. С.Л.Соболева, изд. ООО "Дигит Про", Т.2. 2017. с. 138-147.
2. E.Cherkashin, A.Shigarov, I.Orlova, I.Mikhailov. Authoring and Publishing Text Documents by means of Linked Open Data Technologies. Procs. of International Conference on Applied Internet and Information Technologies (ICAIIT) 2017. Zrenjanin, October 5-6, Serbia, 2017. p. 98-109. URL: <http://tfzr.rs/aiit/files/ProceedingsAIIT2017.pdf> (access date: 18.10.2017) ISBN 978-86-7672-304-1